Download Cost of Private Updating

Bryttany Herren[1], Ahmed Arafa[1], and Karim Banawan[2]

[1]Electrical and Computer Engineering Department, University of North Carolina at Charlotte, USA
[2]Department of Electrical Engineering, Alexandria University, Egypt

***Abstract*—We consider the problem of privately updating a message out of K messages from N replicated and non-colluding databases. In this problem, a user has an outdated version of the message $\hat{W}_\theta$ of length L bits that differ from the current version W_θ in at most f bits. The user needs to retrieve W_θ correctly using a private information retrieval (PIR) scheme with the least number of downloads without leaking any information about the message index θ to any individual database. To that end, we propose a novel achievable scheme based on *syndrome decoding*. Specifically, the user downloads the syndrome corresponding to W_θ, according to a linear block code with carefully designed parameters, using the optimal PIR scheme for messages with a length constraint. We derive lower and upper bounds for the optimal download cost that match if the term $\log_2\left(\sum_{i=0}^{f}\binom{L}{i}\right)$ is an integer. Our results imply that there is a significant reduction in the download cost if $f < \frac{L}{2}$ compared with downloading W_θ directly using classical PIR approaches without taking the correlation between W_θ and $\hat{W}_\theta$ into consideration.**

I. Introduction

The problem of private information retrieval (PIR), introduced by Chor et al. in [1], seeks to find the most efficient way for a user to privately retrieve a single message from a set of K messages from N fully replicated and non-communicating databases. PIR schemes are designed to download a *mixture* of all K messages, with the least number of overhead downloaded bits, such that no single database can infer the identity of the desired message. The user accomplishes this task by sending a query to each database. The databases respond truthfully to the submitted query with an answer string. The user can reconstruct the desired message from jointly *decoding* the returned answer strings. Recently, the problem of PIR has received a growing interest from the information and coding theory communities. The classical PIR problem is re-formulated using information-theoretic measures in the seminal work of Sun-Jafar [2]. In there, the performance metric of the PIR scheme is the retrieval rate, which is the ratio of the number of the desired message symbols to the total number of downloaded bits. The supremum of this ratio is denoted by the PIR capacity, C. Sun and Jafar characterize the PIR capacity of the classical PIR model to be

$$C = \left(1 + \frac{1}{N} + \frac{1}{N^2} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}. \tag{1}$$

Following [2], the capacity (or its reciprocal, the normalized download cost) of many variations of the problem have been investigated, see, e.g., [3]–[17].

In all these works, the user is assumed to have no information about the desired message prior to retrieval. Thus, the queries are designed independently from the message contents. This is not always the case in practice. To see that, consider the following classical motivational example of PIR: in the stock market, investors need to privately retrieve some of the stock records, since showing interest in a specific record may undesirably affect its value. PIR is a natural solution to this problem. Now, consider the case when an investor has already retrieved a specific stock record some time ago but this record has been changed. The investor needs to update the record at his/her side. A trivial solution to this problem is to re-apply the original PIR scheme again. Nevertheless, this solution overlooks the fact that stock records are *correlated* in time. Another example arises in the context of private federated submodel learning [18], in which a user needs to retrieve the up-to-date desired submodel without leaking any information about its identity. The weights of each submodel are usually correlated in time as in the stock market example. In both examples, it is interesting to investigate whether or not the investor (user) can exploit the correlation between the outdated record (submodel) and its up-to-date counterpart to drive down the download cost. In this work, we focus our attention on a specific type of correlation, in which the up-to-date message is a distorted version of the outdated message according to a *Hamming distortion* measure. The most closely related works to this problem are the PIR problems with side information, e.g., [19]–[25]. In all these works, the user has side information in the form of a subset of *undesired* messages, which are utilized to assist in privately retrieving the desired message. This is different from our setting, in which the user possesses side information in the form of an outdated *desired* message. Furthermore, these works differ from each other in whether the privacy of the side information should be maintained or not. This is different from our problem in which the identity of the desired and side information is the same, and therefore the privacy constraint in our problem is modified to reflect this fact.

In this paper, we introduce the problem of *private updating* for a message out of a K-message library from N replicated and non-colluding databases. In this problem, the user has an *outdated* version of the desired message $\hat{W}_\theta$, and wishes to update it to its up-to-date version W_θ. Furthermore, the user has information about the *maximum* Hamming distance f between the up-to-date message and its outdated counterpart, i.e., the user possesses $\hat{W}_\theta$, which differs in *at most* f bits

from the desired up-to-date message W_θ. Based on $\hat{W}_\theta$ and f, the user needs to design a query set to reliably and privately decode the up-to-date version of the desired message W_θ with the least number of downloaded bits. Equivalently, the user needs to privately retrieve an *auxiliary* message that corresponds to the flipped bit positions in the desired message. Similar to the works of [26], [27], we assume that the databases can construct a *mapping* from the original library of messages into a more appropriate form that can assist the user in the retrieval process. We aim at characterizing the optimal download cost needed to update $\hat{W}_\theta$ to W_θ without disclosing the desired message index θ to any of the databases.

To that end, we propose a novel achievable scheme that is based on the *syndrome decoding* idea introduced in [28], and adapt it to our setting to exploit the correlation between W_θ and $\hat{W}_\theta$. Hence, syndrome decoding is used to *compress* the desired message based on the user's side information (i.e., the outdated message $\hat{W}_\theta$). More specifically, the databases apply a linear transformation to the stored library of messages using the parity check matrix of a linear block code with carefully chosen parameters. The existence of such a code can be readily inferred from the Gilbert-Varshamov and the Hamming bounds [29]. This transformation, in effect, maps the messages into their corresponding syndromes. Thus, the problem is reduced to retrieving the auxiliary messages (i.e., the syndrome representation) that comprises of $\lceil \bar{L} \rceil = \left\lceil \log_2 \left(\sum_{i=0}^{f} \binom{L}{i} \right) \right\rceil \leq L$ bits, where L is the original message length. This enables us to directly apply the PIR scheme in [30] to the auxiliary messages of length $\lceil \bar{L} \rceil$, which is optimal under message length constraints. We confirm the validity of our proposed scheme by deriving a matching converse proof. Our converse proof is inspired by the converse proofs of the PIR problem with side information in [19], [20], with the main difference being the fact that the side information in our case is the outdated message $\hat{W}_\theta$ in contrast to the cached messages. Consequently, we show that the optimal download cost, $\bar{D}_L$, is bounded by $\left\lceil \frac{\bar{L}}{C} \right\rceil \leq \bar{D}_L \leq \left\lceil \frac{\lceil \bar{L} \rceil}{C} \right\rceil$. Our achievable scheme is optimal if $\bar{L}$ is an integer, otherwise the gap between the upper and lower bounds is upper bounded by 2 bits. This justifies the efficacy of using syndromes as a message mixing technique in our setting. Furthermore, our results show that performing direct PIR on the original library of messages is strictly sub-optimal as long as the maximum Hamming distance $f < \frac{L}{2}$.

II. System Model

We consider a classical PIR problem with K independent, uncoded, messages $W_1, \cdots, W_K$, with each message consisting of L independent and uniformly distributed bits. We have

$$H(W_i) = L, \quad 1 \leq i \leq K, \tag{2}$$

$$H(W_1, \cdots, W_K) = H(W_1) + \cdots + H(W_K). \tag{3}$$

The K messages are stored in N replicated and non-communicating databases. The user (retriever) has a local copy of one of the messages whose index $\theta \in [K]$ is known to the user,[1] but not the database.[2] However, this message stored locally is *outdated*, and the user wishes to update it so that it is consistent with the copies in the databases without revealing to any of the databases what the message index is. This setting defines the *private updating problem*.

Since each message is a string of L bits, the problem can be formulated as privately determining which subset of the message bits need to be flipped in order to fully update it. To model this, we use $\hat{W}_\theta$ to represent the locally stored outdated message, $\bar{W}_\theta$ to represent the subset of bit indices that need to be flipped, and f to represent the *maximum* Hamming distance between W_θ and $\hat{W}_\theta$. Therefore, in order to update message θ the user needs to flip *at most* f bits, i.e., $\bar{W}_\theta$ takes a value out of $\sum_{i=0}^{f} \binom{L}{i}$ choices. We assume that such choices are uniformly distributed and independently realized from $\hat{W}_\theta$. Based on this model, the following holds:

$$H(W_\theta) = H(\hat{W}_\theta) = L, \tag{4}$$

$$H(\bar{W}_\theta) = \log_2 \left(\sum_{i=0}^{f} \binom{L}{i} \right) \triangleq \bar{L}, \tag{5}$$

$$H(W_\theta | \hat{W}_\theta) = H(\bar{W}_\theta | \hat{W}_\theta) = \bar{L}, \tag{6}$$

$$H(\bar{W}_\theta | \hat{W}_\theta, W_\theta) = 0, \tag{7}$$

$$|\bar{W}_\theta| \leq f \leq L, \tag{8}$$

where $|\cdot|$ denotes cardinality. For the purposes of this paper, we assume that the maximum Hamming distance f between the outdated and updated message is known to the user.

In order to retrieve W_θ, the user sends a set of queries $Q_1^{[\hat{W}_\theta, f]}, \ldots, Q_N^{[\hat{W}_\theta, f]}$ to the N databases to efficiently obtain $\bar{W}_\theta$. The queries are generated according to $\hat{W}_\theta$ and f, and are jointly independent of the realizations of the $[K] \backslash \{\theta\}$ messages and $\bar{W}_\theta$ given $\hat{W}_\theta$. Therefore we have[3]

$$I\left(W_{[K]\backslash\{\theta\}}, \bar{W}_\theta; Q_{1:N}^{[\hat{W}_\theta, f]} \middle| \hat{W}_\theta \right) = 0. \tag{9}$$

Upon receiving the query $Q_n^{[\hat{W}_\theta, f]}$, the nth database replies with an answering string $A_n^{[\hat{W}_\theta, f]}$, which is a function of $Q_n^{[\hat{W}_\theta, f]}$ and all the K messages stored. Therefore, $\forall \theta \in [K],\ \forall n \in [N]$, we have

$$H\left(A_n^{[\hat{W}_\theta, f]} \middle| Q_n^{[\hat{W}_\theta, f]}, W_{1:K} \right) = 0. \tag{10}$$

To ensure that individual databases do not know which message is being updated, we need to satisfy the following *privacy constraint*, $\forall n \in [N],\ \forall k \in [K]$:

$$\left(Q_n^{[\hat{W}_1, f]}, A_n^{[\hat{W}_1, f]}, \hat{W}_1, W_{1:K} \right) \sim \left(Q_n^{[\hat{W}_k, f]}, A_n^{[\hat{W}_k, f]}, \hat{W}_k, W_{1:K} \right), \tag{11}$$

where $\sim$ denotes statistical equivalence. After receiving the answering strings $A_{1:N}^{[\hat{W}_\theta, f]}$ from all the N databases, the

[1] $[K]$ denotes the set $\{1, 2, \ldots, K\}$.

[2] This is true if message θ has been previously obtained in a private manner.

[3] We use the notation x_S to denote the collection of $\{x_i,\ i \in S\}$.

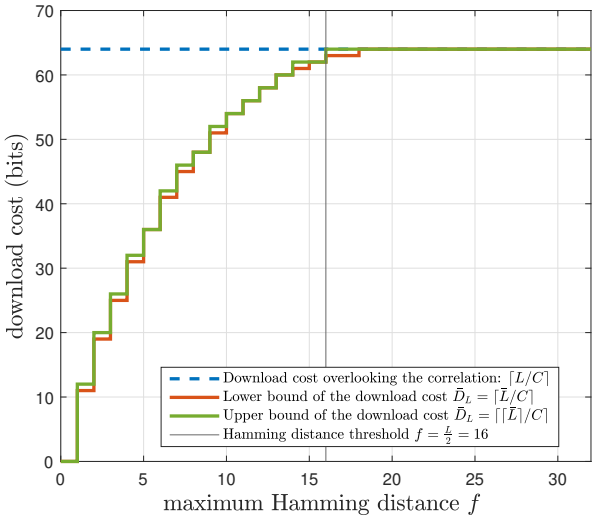

Fig. 1: Download cost of private updating with $L = 32$ bits, $N = 2$ databases, and $K = 10$ messages.

user needs to decode the desired information W_θ with no uncertainty, satisfying the following *correctness constraint*:

$$H\left(W_\theta \middle| A_{1:N}^{[\hat{W}_\theta, f]}, Q_{1:N}^{[\hat{W}_\theta, f]}, \hat{W}_\theta\right) = 0. \tag{12}$$

For fixed N, K, and f, a pair $(\bar{D}, L)$ is *achievable* if there exists a private updating scheme for messages of length L bits long satisfying the privacy constraint (11) and the correctness constraint (12). In this pair, $\bar{D}$ represents the expected number of downloaded bits received from the N databases independently via the answering strings $A_{1:N}^{[\hat{W}_k, f]}$, i.e.,

$$\bar{D} = \sum_{n=1}^{N} H\left(A_n^{[\hat{W}_\theta, f]}\right). \tag{13}$$

Our goal is to characterize the optimal download cost $\bar{D}_L$ for fixed arbitrary N, K, and f. That is, to solve for

$$\bar{D}_L = \min\left\{\bar{D} : (\bar{D}, L) \text{ is achievable}\right\}. \tag{14}$$

Clearly, the user can ignore its outdated message $\hat{W}_\theta$ and re-download the whole new message W_θ using standard PIR schemes [2]. In the next section, however, we show that we can use $\hat{W}_\theta$ to do strictly better.

III. Main Result

We present our main result in the following theorem:

Theorem 1 *In the private updating problem, we have*

$$\left\lceil \frac{\bar{L}}{C} \right\rceil \leq \bar{D}_L \leq \left\lceil \frac{\lceil \bar{L} \rceil}{C} \right\rceil, \tag{15}$$

with C and $\bar{L}$ defined in (1) and (5), respectively.

Fig. 1 shows the efficiency of our result by plotting the upper and lower bounds of the download cost for the private updating problem with $L = 32$ bits, $N = 2$ databases, and $K = 10$ messages.

We show the first inequality in (15) by presenting a converse proof for Theorem 1 in Section IV, which is based on similar arguments to those used in cache-aided PIR settings [20]. The second inequality in (15) is shown by a novel achievability scheme for Theorem 1 in Section V, which is based on distributed source coding [28]. We now have some remarks.

Remark 1 *From* (5) *and* (8)*, it follows that $\lceil \bar{L} \rceil = L$ for all values of $f \geq \frac{L}{2}$; and that $\lceil \bar{L} \rceil < L$ for all values of $f < \frac{L}{2}$.*[4] *This means that there is a* Hamming distance threshold *of $\frac{L}{2}$ beyond which there is no advantage to using a private updating strategy, and below which there will always be some savings in download cost (see Fig. 1).*

Remark 2 *If L and f are such that $\bar{L} = \lceil \bar{L} \rceil$ then the two bounds in Theorem 1 match. We will see that this holds if a* perfect code[5] *by which the queries are sent exists (cf. Section V). Otherwise, if $\bar{L} < \lceil \bar{L} \rceil$, one can show that the two bounds are within 2 bits for $N \geq 2$ databases (see [30, Section 7.2]).*

IV. Proof of Main Result: Converse

In this section, we show that $\lceil \bar{L}/C \rceil$ serves as a general lower bound for the download cost in (13). To do so, we prove two useful lemmas, which were previously used in the cache-aided PIR setting of [20], for the case of our private updating problem. The two lemmas are then combined to prove the general lower bound. The key difference between our lemmas and those in [20] is that rather than a set of cached messages, the user is given an outdated message $\hat{W}_\theta$, requiring careful handling of the correlation between W_θ and $\hat{W}_\theta$. Without loss of generality, we re-label the messages such that $\theta = 1$.

Lemma 1 (Interference lower bound) *In the private updating problem, the interference from undesired messages within the answering strings, $\bar{D} - \bar{L}$, satisfies*

$$\bar{D} - \bar{L} \geq I\left(W_{2:K}; Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]} \middle| W_1, \hat{W}_1\right). \tag{16}$$

Proof: We start with the right hand side of (16),

$$\begin{aligned}
&I(W_{2:K}; Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]} | W_1, \hat{W}_1) \\
&= I(W_{2:K}; Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]}, W_1 | \hat{W}_1) \\
&\quad - I(W_{2:K}; W_1 | \hat{W}_1) && (17) \\
&= I(W_{2:K}; Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]} | \hat{W}_1) \\
&\quad + I(W_{2:K}; W_1 | Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]}, \hat{W}_1) && (18) \\
&\overset{(12)}{=} I(W_{2:K}; Q_{1:N}^{[\hat{W}_1, f]}, A_{1:N}^{[\hat{W}_1, f]} | \hat{W}_1) && (19) \\
&\overset{(9)}{=} I(W_{2:K}; A_{1:N}^{[\hat{W}_1, f]} | Q_{1:N}^{[\hat{W}_1, f]}, \hat{W}_1) && (20) \\
&= H(A_{1:N}^{[\hat{W}_1, f]} | Q_{1:N}^{[\hat{W}_1, f]}, \hat{W}_1) \\
&\quad - H(A_{1:N}^{[\hat{W}_1, f]} | Q_{1:N}^{[\hat{W}_1, f]}, W_{2:K}, \hat{W}_1) && (21)
\end{aligned}$$

[4] This can be readily shown using the binomial theorem. Details are omitted.
[5] Perfect codes are those that attain the Hamming bound with equality [29].

$$\overset{(12)}{=} H(A_{1:N}^{[\hat{W}_1,f]}|Q_{1:N}^{[\hat{W}_1,f]},\hat{W}_1) - H(A_{1:N}^{[\hat{W}_1,f]},W_1|Q_{1:N}^{[\hat{W}_1,f]},W_{2:K},\hat{W}_1) \quad (22)$$

$$\leq H(A_{1:N}^{[\hat{W}_1,f]}|Q_{1:N}^{[\hat{W}_1,f]},\hat{W}_1) - H(W_1|Q_{1:N}^{[\hat{W}_1,f]},W_{2:K},\hat{W}_1) \quad (23)$$

$$\overset{(9)}{=} H(A_{1:N}^{[\hat{W}_1,f]}|Q_{1:N}^{[\hat{W}_1,f]},\hat{W}_1) - H(W_1|W_{2:K},\hat{W}_1) \quad (24)$$

$$\overset{(13),(6)}{\leq} \bar{D} - \bar{L}. \quad (25)$$

This concludes the proof. ■

Note that if privacy was not a constraint, then $\bar{D} = \bar{L}$ and the interference from undesired messages would be non-existent. However, when the privacy constraint is present, $\bar{D}-\bar{L}$ characterizes the number of bits that will be downloaded and used as side information to preserve privacy from the databases in a given scheme.

Lemma 2 (Induction lemma) *For all $k \in \{2,\dots,K\}$, the mutual information term in Lemma 1 can be inductively lower bounded as*

$$I\left(W_{k:K};Q_{1:N}^{[\hat{W}_{k-1},f]},A_{1:N}^{[\hat{W}_{k-1},f]}\middle|W_{1:k-1},\hat{W}_{k-1}\right) \geq \frac{1}{N}I\left(W_{k+1:K};Q_{1:N}^{[\hat{W}_k,f]},A_{1:N}^{[\hat{W}_k,f]}\middle|W_{1:k},\hat{W}_k\right)+\frac{\bar{L}}{N}. \quad (26)$$

Proof: We start with the left hand side of (26),

$$I(W_{k:K};Q_{1:N}^{[\hat{W}_{k-1},f]},A_{1:N}^{[\hat{W}_{k-1},f]}|W_{1:k-1},\hat{W}_{k-1}) \geq \frac{1}{N}\sum_{n=1}^{N} I(W_{k:K};Q_n^{[\hat{W}_{k-1},f]},A_n^{[\hat{W}_{k-1},f]}|W_{1:k-1},\hat{W}_{k-1}) \quad (27)$$

$$\overset{(11)}{=} \frac{1}{N}\sum_{n=1}^{N} I(W_{k:K};Q_n^{[\hat{W}_k,f]},A_n^{[\hat{W}_k,f]}|W_{1:k-1},\hat{W}_k) \quad (28)$$

$$\overset{(9)}{=} \frac{1}{N}\sum_{n=1}^{N} I(W_{k:K};A_n^{[\hat{W}_k,f]}|W_{1:k-1},\hat{W}_k,Q_n^{[\hat{W}_k,f]}) \quad (29)$$

$$\overset{(10)}{=} \frac{1}{N}\sum_{n=1}^{N} H(A_n^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k,Q_n^{[\hat{W}_1,f]}) \quad (30)$$

$$\geq \frac{1}{N}\sum_{n=1}^{N} H(A_n^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k,Q_{1:N}^{[\hat{W}_1,f]},A_{1:n-1}^{[\hat{W}_1,f]}) \quad (31)$$

$$\overset{(10)}{=} \frac{1}{N}\sum_{n=1}^{N} I(W_{k:K};A_n^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k,Q_{1:N}^{[\hat{W}_1,f]},A_{1:n-1}^{[\hat{W}_1,f]}) \quad (32)$$

$$= \frac{1}{N}I(W_{k:K};A_{1:N}^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k,Q_{1:N}^{[\hat{W}_1,f]}) \quad (33)$$

$$\overset{(9)}{=} \frac{1}{N}I(W_{k:K};Q_{1:N}^{[\hat{W}_1,f]},A_{1:N}^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k) \quad (34)$$

$$\overset{(12)}{=} \frac{1}{N}I(W_{k:K};W_k,Q_{1:N}^{[\hat{W}_1,f]},A_{1:N}^{[\hat{W}_1,f]}|W_{1:k-1},\hat{W}_k) \quad (35)$$

$$= \frac{1}{N}I(W_{k:K};Q_{1:N}^{[\hat{W}_1,f]},A_{1:N}^{[\hat{W}_1,f]}|W_{1:k},\hat{W}_k) + \frac{1}{N}I(W_{k:K};W_k|W_{1:k-1},\hat{W}_k) \quad (36)$$

$$\overset{(6)}{=} \frac{1}{N}I(W_{k+1:K};Q_{1:N}^{[\hat{W}_1,f]},A_{1:N}^{[\hat{W}_1,f]}|W_{1:k},\hat{W}_1)+\frac{\bar{L}}{N}. \quad (37)$$

This concludes the proof. ■

We now apply the result of Lemma 2 recursively on that of Lemma 1 to get the general lower bound.

Lemma 3 *The optimal private updating download cost satisfies the following lower bound:*

$$\bar{D}_L \geq \left\lceil \bar{L}\left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}}\right)\right\rceil = \left\lceil \frac{\bar{L}}{C}\right\rceil. \quad (38)$$

Proof: Any private updating scheme's download cost satisfies the following series of inequalities:

$$\bar{D} \overset{(16)}{\geq} \bar{L} + I(W_{2:K};Q_{1:N}^{[\hat{W}_1,f]},A_{1:N}^{[\hat{W}_1,f]}|W_1,\hat{W}_1) \quad (39)$$

$$\overset{(26)}{\geq} \bar{L}+\frac{\bar{L}}{N}+\frac{1}{N}I(W_{3:K};Q_{1:N}^{[\hat{W}_2,f]},A_{1:N}^{[\hat{W}_2,f]}|W_{1:2},\hat{W}_2) \quad (40)$$

$$\overset{(26)}{\geq} \bar{L}+\frac{\bar{L}}{N}+\frac{\bar{L}}{N^2} + \frac{1}{N}I(W_{4:K};Q_{1:N}^{[\hat{W}_3,f]},A_{1:N}^{[\hat{W}_3,f]}|W_{1:3},\hat{W}_3) \quad (41)$$

$$\overset{(26)}{\geq} \cdots \quad (42)$$

$$\overset{(26)}{\geq} \bar{L}\left(1+\frac{1}{N}+\cdots+\frac{1}{N^{K-1}}\right) \overset{(1)}{=} \frac{\bar{L}}{C}. \quad (43)$$

Since (43) lower bounds the download cost $\bar{D}$ for *any* private updating scheme, it also lower bounds the download cost of the *optimal* private updating scheme $\bar{D}_L$. Finally, since $\bar{D}_L$ is an integer, we take the ceiling of (43) to get (38). ■

This shows that the first inequality of (15) holds, and concludes the converse proof.

V. Proof of the Main Result: Achievability

Our achievability scheme makes use of the correlation between W_θ and $\hat{W}_\theta$ through the knowledge of their maximum Hamming distance f in order to reduce the download cost. This approach is related to the problem tackled in [28] (without privacy constraints), in which a source is compressed given that it is correlated with some side information that is available only at the decoder. The retrieving user represents the decoder in our case, with side information $\hat{W}_\theta$. By the Slepian-Wolf coding theorem [31], one can noiselessly compress the source W_θ at the rate of $H(W_\theta|\hat{W}_\theta) = \bar{L}$. The *compressed* source is treated as a *new message* to be downloaded using a PIR scheme, as opposed to downloading the whole message W_θ. Such scheme, however, has a message length constraint (unlike most of the PIR works in the literature). For that reason, we leverage tools from the PIR scheme with arbitrary message length in [30] to accomplish our task. The details are illustrated in the motivating examples below.

A. Example: $N = 2$, $K = 2$, $L = 3$, and $f = 1$

In this example, we have $\bar{L} = \log_2(1+3) = 2$, and $C = 2/3$. We show that $\bar{D} = \lceil \lceil \bar{L} \rceil / C \rceil = 3$ bits is achievable. We

first start by constructing a $[3,1,3]$ linear block code, which is in this case a repetition code with generator matrix G and parity check matrix H given by

$$\mathsf{G} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \quad \mathsf{H} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{44}$$

Note that such code is capable of correcting at most $f = 1$ error. The syndromes associated with this code are $\mathsf{s} \in \{00, 01, 10, 11\}$. Observe that the length of s is exactly $\lceil \bar{L} \rceil$.

Instead of requesting W_θ, the user retrieves the index of the coset in which W_θ resides in the code's standard array. That is, its corresponding syndrome

$$\mathsf{s}_\theta = W_\theta \mathsf{H}^T. \tag{45}$$

The user then compares $\hat{W}_\theta$ to all the words in that coset, and decodes W_θ as the one closest in Hamming distance. This is guaranteed to yield the unique correct message [28]. Therefore, the syndrome s_θ efficiently represents the flipped bits' indices $\bar{W}_\theta$, and one is able to reduce the effective message length from $L = 3$ to $\lceil \bar{L} \rceil = 2$ by dealing with the syndrome s_θ instead of W_θ.

Let $W_1 = [a_1, a_2, a_3]$, and $W_2 = [b_1, b_2, b_3]$. The syndromes (the new messages) are given by

$$\mathsf{s}_1 = W_1 \mathsf{H}^T = \begin{bmatrix} a_1 + a_2 & a_1 + a_3 \end{bmatrix} \triangleq \begin{bmatrix} \bar{a}_1 & \bar{a}_2 \end{bmatrix}, \tag{46}$$

$$\mathsf{s}_2 = W_2 \mathsf{H}^T = \begin{bmatrix} b_1 + b_2 & b_1 + b_3 \end{bmatrix} \triangleq \begin{bmatrix} \bar{b}_1 & \bar{b}_2 \end{bmatrix}. \tag{47}$$

Assume $\theta = 1$. Since $\lceil \bar{L} \rceil = N^{K-1}$, we can apply a *non-symmetric* PIR scheme as follows to decode s_1 [30]:

Database 1	Database 2
$\bar{a}_1$, $\bar{b}_1$	$\bar{a}_2 + \bar{b}_1$

This has a download cost of $\bar{D} = 3$ bits, which is optimal in this case since it meets the converse bound.

The repetition code used in this example is a *perfect code.* While this makes $\bar{L}$ an integer, and meets the converse bound, perfect codes are scarce. In the next example, we show how the proposed scheme performs with non-perfect codes.

B. Example: N=2, K=2, L=5, and f=1

In this example, we have $\bar{L} = \log_2(1+5) = 2.58$, and $C = 2/3$. We show that $\bar{D} = \left\lceil \lceil \bar{L} \rceil / C \right\rceil = 5$ bits is achievable. As in the previous example, we start by constructing a $[5,2,3]$ linear block code. Differently though, this is not a repetition code, and is characterized by

$$\mathsf{G} = \begin{bmatrix} 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad \mathsf{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \tag{48}$$

The syndromes s have length $\lceil \bar{L} \rceil$. Specifically,

$$\mathsf{s}_1 = W_1 \mathsf{H}^T = \begin{bmatrix} a_1 + a_2 + a_3 & a_1 + a_2 + a_4 & a_2 + a_5 \end{bmatrix} \triangleq \begin{bmatrix} \bar{a}_1 & \bar{a}_2 & \bar{a}_3 \end{bmatrix}, \tag{49}$$

$$\mathsf{s}_2 = W_2 \mathsf{H}^T = \begin{bmatrix} b_1 + b_2 + b_3 & b_1 + b_2 + b_4 & b_2 + b_5 \end{bmatrix} \triangleq \begin{bmatrix} \bar{b}_1 & \bar{b}_2 & \bar{b}_3 \end{bmatrix}. \tag{50}$$

Since $\lceil \bar{L} \rceil = N^{K-1} + 1$, we follow the methodology in [30]; we privately download $N^{K-1} = 2$ bits ($\bar{a}_1$ and $\bar{a}_2$) using the non-symmetric PIR scheme in the previous example, and then privately download the remaining 1 bit ($\bar{a}_3$) using the scheme in [32]. The technique in [32] in this case is such that the user requests random linear combinations of $[\bar{a}_3 \ \bar{b}_3]$ from database 1 using a random binary vector $\boldsymbol{h}$, and the same from database 2 yet with $\boldsymbol{h}' = \boldsymbol{h} + \boldsymbol{e}_\theta$, where $\boldsymbol{e}_i$ is the ith standard basis vector. The full PIR scheme is as follows:

Database 1	Database 2
$\bar{a}_1$, $\bar{b}_1$	$\bar{a}_2 + \bar{b}_1$
$h_1 \bar{a}_3 + h_2 \bar{b}_3$	$(h_1+1)\bar{a}_3 + h_2 \bar{b}_3$

This has a download cost of $\bar{D} = 5$ bits, which is 1 bit away from the converse bound since the code used is non-perfect.

C. The General Scheme

For general N, K, L, and f, we construct an $[L, L - \lceil \bar{L} \rceil, 2f+1]$ linear block code. From the Gilbert-Varshamov bound [29], we know that such a code exists if

$$2^{\lceil \bar{L} \rceil} \leq \sum_{j=0}^{2f} \binom{L}{j}. \tag{51}$$

In addition, such a code must satisfy the Hamming bound [29]:

$$\sum_{j=0}^{f} \binom{L}{j} \leq 2^{\lceil \bar{L} \rceil}. \tag{52}$$

By the definition of $\bar{L}$ in (5), both (51) and (52) are satisfied, and so the code exists and is able to correct f bit flips.

Next, we map each message to its corresponding syndrome of the constructed code, which is of length $L - (L - \lceil \bar{L} \rceil) = \lceil \bar{L} \rceil$. The user then retrieves the syndrome s_θ according to a PIR scheme with N databases, K messages, and $\lceil \bar{L} \rceil$ message length. By [30, Theorem 1], a download cost of $\left\lceil \lceil \bar{L} \rceil / C \right\rceil$ is achievable in this case. Finally, correctness is guaranteed since querying for the syndrome s_θ allows the user to decode W_θ as the unique word in the syndrome's coset with the least Hamming distance from $\hat{W}_\theta$ [28].

This shows that the second inequality of (15) holds, and concludes the achievability proof.

VI. Conclusions and Discussions

In this work, a novel private updating problem has been introduced, in which a user's outdated message is to be privately updated by querying a set of replicated and non-colluding databases that have the up-to-date version. Under a Hamming distortion measure between the outdated and the up-to-date messages, a syndrome decoding technique is leveraged to compress the number of bits that needs to be downloaded in order to correctly update the message. This has been combined with PIR schemes with message length constraints to guarantee privacy. The proposed private updating scheme has been shown to be optimal when the system parameters enable the construction of a perfect code according to which

the syndrome decoding technique is worked out. In other cases, the achievable download cost has been shown to be within at most 2 bits from a derived converse bound.

The model of this paper assumes that the Hamming distortion between W_θ and $\hat{W}_\theta$ is upper bounded by f. If, instead, the Hamming distortion is *known to be exactly* f', then the download cost can be reduced, and using codes to map the messages into syndromes may be unnecessary. To see this, consider an example with $N = 2$, $K = 2$, $L = 8$, and $f = 1$. In this case, $\lceil \bar{L} \rceil = 4$, and by Theorem 1, a download cost of 6 bits is achievable.

Let us now set $f' = 1$, i.e., the user knows that W_θ and $\hat{W}_\theta$ differ in *exactly* 1 bit. Assuming $\theta = 1$, define $\bar{a}_1 \triangleq a_1 + a_2 + a_3 + a_4$, $\bar{a}_2 \triangleq a_1 + a_2 + a_5 + a_6$, and $\bar{a}_3 \triangleq a_1 + a_3 + a_5 + a_7$, where a_i's represent the bits of the desired message W_1. The user then constitutes similar combinations using the bits of the outdated message $\hat{W}_1$ to get $\hat{a}_1$, $\hat{a}_2$, and $\hat{a}_3$. Now observe that possessing the *new message* $[\bar{a}_1, \bar{a}_2, \bar{a}_3]$ is sufficient to determine the position of the flipped bit by comparing it to $[\hat{a}_1, \hat{a}_2, \hat{a}_3]$. For instance, if $\bar{a}_i = \hat{a}_i,\ \forall i$, then $\hat{W}_1(8)$ needs to be flipped. If on the other hand $\bar{a}_i \neq \hat{a}_i,\ \forall i$, then $\hat{W}_1(1)$ needs to be flipped. While if $\bar{a}_1 \neq \hat{a}_1$, and $\bar{a}_i = \hat{a}_i,\ i = 2, 3$, then $\hat{W}_1(4)$ needs to be flipped, and so on. Therefore, the effective message length is reduced to 3 (as opposed to $\lceil \bar{L} \rceil = 4$), and a download cost of 5 bits is achievable.

The above procedure can be done using a *bisection search* approach. The user can first retrieve $a_1 + a_2 + a_3 + a_4$, and compares it to the sum of the outdated message's first 4 bits. If they are equal, then the error must lie in the last 4 bits of $\hat{W}_1$. Assuming this is the case, the user downloads $a_5 + a_6$, and compares it to $\hat{W}_1(5) + \hat{W}_1(6)$. If they too are equal, then the error must lie in the last 2 bits of $\hat{W}_1$. Assuming this is the case as well, the user finally downloads a_7 and compares it to $\hat{W}_1(7)$. If they are equal, then $\hat{W}_1(8)$ needs to be flipped. We see that this bisection approach has an effective message length of $\log_2(L) = 3$ bits. However, since the next query structure depends on the answers of the previous queries, a *multiround* PIR scheme needs to devised in this case [33].

It would be interesting to extend the results of this paper to work for the case of known distortion f' (and generally for other notions of correlation measures between W_θ and $\hat{W}_\theta$), which may be relevant in certain applications.

REFERENCES

[1] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, p. 965–981, Nov. 1998.

[2] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, pp. 4075–4088, July 2017.

[3] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, March 2018.

[4] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Transactions on Information Theory*, vol. 65, pp. 322–329, January 2019.

[5] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. on Info. Theory*, vol. 64, pp. 6842–6862, October 2018.

[6] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, C. Hollanti, and S. E. Rouayheb, "Private information retrieval schemes for coded data with arbitrary collusion patterns," in *Proc. IEEE ISIT*, June 2017.

[7] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, pp. 3183–3197, May 2019.

[8] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, pp. 7613–7627, November 2019.

[9] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2999–3012, March 2020.

[10] K. Banawan and S. Ulukus, "The capacity of private information retrieval from byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, pp. 1206–1219, February 2019.

[11] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," *IEEE Trans. Inf. Theory*, vol. 66, pp. 6617–6634, November 2020.

[12] H. Sun and S. A. Jafar, "The capacity of private computation," *IEEE Trans. Inf. Theory*, vol. 65, pp. 3880–3897, June 2019.

[13] S. Kumar, A. G. i Amat, E. Rosnes, and L. Senigagliesi, "Private information retrieval from a cellular network with caching at the edge," *IEEE Trans. Commun.*, vol. 67, pp. 4900–4912, July 2019.

[14] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Trans. Inf. Theory*, vol. 66, pp. 3590–3602, June 2020.

[15] X. Yao, N. Liu, and W. Kang, "The capacity of multi-round private information retrieval from Byzantine databases," in *Proc. IEEE ISIT*, July 2019.

[16] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *Proc. IEEE ISIT*, July 2019.

[17] R. G. L. D'Oliveira and S. El Rouayheb, "One-shot PIR: Refinement and lifting," *IEEE Trans. Inf. Theory*, vol. 66, pp. 2443–2455, April 2020.

[18] Z. Jia and S. Jafar, "X-secure T-private federated submodel learning," [Online]. Available: arXiv:2010.01059.

[19] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Trans. Inf. Theory*, vol. 66, pp. 4761–4773, August 2020.

[20] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Inf. Theory*, vol. 65, pp. 8222–8231, December 2019.

[21] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, 2019. Early Access.

[22] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," in *Proc. IEEE ITW*, November 2018.

[23] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *Proc. Allerton*, October 2018.

[24] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *Proc. Allerton*, October 2018.

[25] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, pp. 2032–2043, April 2020.

[26] Z. Chen, Z. Wang, and S. A. Jafar, "The asymptotic capacity of private search," *IEEE Trans. Inf. Theory*, vol. 66, pp. 4709–4721, August 2020.

[27] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," [Online]. Available: arXiv:1912.13501.

[28] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS): design and construction," *IEEE Trans. Inf. Theory*, vol. 49, pp. 626–643, March 2003.

[29] R. E. Blahut, *Algebraic codes for data transmission*. Cambridge university press, 2003.

[30] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, pp. 2920–2932, December 2017.

[31] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. IT-19, pp. 471–480, July 1973.

[32] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *Proc. IEEE ISIT*, June 2014.

[33] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, pp. 5743–5754, August 2018.